Using N-Grams to Understand the Nature of Summaries

Michele Banko and Lucy Vanderwende
One Microsoft Way
Redmond, WA 98052
{mbanko, lucyv}@microsoft.com

Abstract

Although single-document summarization is a well-studied task, the nature of multi-document summarization is only beginning to be studied in detail. While close attention has been paid to what technologies are necessary when moving from single to multi-document summarization, the properties of human-written multi-document summaries have not been quantified. In this paper, we empirically characterize human-written summaries provided in a widely used summarization corpus by attempting to answer the questions: Can multi-document summaries that are written by humans be characterized as extractive or generative? Are multi-document summaries less extractive than single-document summaries? Our results suggest that extraction-based techniques which have been successful for single-document summarization may not be sufficient when summarizing multiple documents.

1 Introduction

The explosion of available online text has made it necessary to be able to present information in a succinct, navigable manner. The increased accessibility of worldwide online news sources and the continually expanding size of the worldwide web place demands on users attempting to wade through vast amounts of text. Document clustering and multi-document summarization technologies working in tandem promise to ease some of the burden on users when browsing related documents.

Summarizing a set of documents brings about challenges that are not present when summarizing a single document. One might expect that a good multi-document summary will present a synthesis of multiple views of the event being described over different documents, or present a high-level view of an event that is not explicitly reflected in any single document. A useful multi-document summary may also indicate the presence of new or distinct information contained within a set of documents describing the same topic (McKeown et. al., 1999, Mani and Bloedorn, 1999). To meet these expectations, a multi-document summary is required to generalize, condense and merge information coming from multiple sources.

Although single-document summarization is a well-studied task (see Mani and Maybury, 1999 for an overview), multi-document summarization is only recently being studied closely (Marcu & Gerber 2001). While close attention has been paid to multi-document summarization technologies (Barzilay et al. 2002, Goldstein et al 2000), the inherent properties of human-written multi-document summaries have not yet been quantified. In this paper, we seek to empirically characterize ideal multi-document summaries in part by attempting to answer the questions: Can multi-document summaries that are written by humans be characterized as extractive or generative? Are multi-document summaries less extractive than single-document summaries? Our aim in answering these questions is to discover how the nature of multi-document summaries will impact our system requirements.

We have chosen to focus our experiments on the data provided for summarization evaluation during the Document Understanding Conference (DUC). While we recognize that other summarization corpora may exhibit different properties than what we report, the data prepared for DUC evaluations is widely used, and continues to be a powerful force in shaping directions in summarization research and evaluation.

In the following section we describe previous work related to investigating the potential for extractive summaries. Section 3 describes a new approach for assessing the degree to which a summary can be described as extractive, and reports our findings for both single and multiple document summarization tasks. We conclude with a discussion of our findings in Section 4.

2 Related Work

Jing (2002) previously examined the degree to which single-document summaries can be characterized as extractive. Based on a manual inspection of 15 human-written summaries, she proposes that for the task of single-document summarization, human summarizers use a "cut-and-paste" approach in which six main operations are performed: sentence reduction, sentence combination, syntactic transformation, reordering, lexical paraphrasing, and generalization or specification. The first four operations are reflected in the construction of an HMM model that can be used to decompose human summaries. According to this model, 81% of summary sentences contained in a corpus of 300 human-written summaries of news articles on telecommunications were found to fit the cut-and-paste method, with the rest believed to have been composed from scratch.[1]

Another recent study (Lin and Hovy, 2003) investigated the extent to which extractive methods may be sufficient for summarization in the single-document case. By computing a performance upper-bound for pure sentence extraction, they found that state-of-the-art extraction-based systems are still 15%-24%[2] away from this limit, and 10% away from average human performance. While this sheds light on how much gain can be achieved by optimizing sentence extraction methods for single-document summarization, to our knowledge, no one has assessed the potential for extraction-based systems when attempting to summarize multiple documents.

3 Using N-gram Sequences to Characterize Summaries

Our approach to characterizing summaries is much simpler than what Jing has described and is based on the following idea: if human-written summaries are extractive, then we should expect to see long spans of text that have been lifted from the source documents to form a summary.

Note that this holds under the assumptions made by Jing's model of operations that are performed by human summarizers. In the examples of operations given by Jing, we notice that long n-grams are preserved (designated by brackets), even in the operations mostly likely to disrupt the original text:

Sentence Reduction:

Document sentence: When it arrives sometime next year in new TV sets, the V-chip will give parents a new and potentially revolutionary device to block out programs they don't want their children to see.

Summary sentence: [The V-chip will give parents a] [device to block out programs they don't want their children to see.]

Syntactic Transformation:

Document sentence: Since annoy.com enables visitors to send unvarnished opinions to political and other figures in the news, the company was concerned that its activities would be banned by the statute.

Summary sentence: [Annoy.com enables visitors to send unvarnished opinions to political and other figures in the news] and feared the law could put them out of business.

Sentence Combination:

Document sentence 1: But it also raises serious questions about the privacy of such highly personal information wafting about the digital world.

Document sentence 2: The issue thus fits squarely into the broader debate about privacy and security on the Internet, whether it involves protecting credit card numbers or keeping children from offensive information.

Summary sentence: [But it also raises] the issue of [privacy of such] [personal information] and this issue hits the nail on the head [in the broader debate about privacy and security on the Internet.]

3.1 Data and Experiments

For our experiments we used data made available from the 2001 Document Understanding Conference (DUC), an annual large-scale evaluation of summarization systems sponsored by the National Institute of Standards and Technology (NIST). In this corpus, NIST has gathered documents describing 60 events, taken from the Associated Press, Wall Street Journal, FBIS San Jose Mercury, and LA Times newswires. An event is described by between 3 and 20 separate (but not necessarily unique) documents; on average a cluster contains 10 documents. Of the 60 available clusters, we used the portion specifically designated for training, which contains a total of 295 documents distributed over 30 clusters.

As part of the DUC 2001 summarization corpus, NIST also provides four hand-written summaries of different lengths for every document cluster, as well as 100-word summaries of each document. Since we wished to collectively compare single-document summaries against multi-document summaries, we used the 100-word multi-document summaries for our analysis. It is important to note that for each cluster, all summaries (50, 100, 200 and 400-word multi-document

[1] Jing considers a sentence to have been generated from scratch if fewer than half of the words were composed of terms coming from the original document.

[2] The range in potential gains is due to possible variations in summary length.

and 100-word per-document) have been written by the same author. NIST used a total of ten authors, each providing summaries for 3 of the 30 topics. The instructions provided did not differ per task; in both single and multi-document scenarios, the authors were directed to use complete sentences and told to feel free to use their own words (Over, 2004).

To compare the text of human-authored multi-document summaries to the full-text documents describing the events, we automatically broke the documents into sentences, and constructed a minimal tiling of each summary sentence. Specifically, for each sentence in the summary, we searched for all n-grams that are present in both the summary and the documents, placing no restrictions on the potential size of an n-gram. We then covered each summary sentence with the n-grams, optimizing to use as few n-grams as possible (i.e. favoring n-grams that are longer in length). For this experiment, we normalized the data by converting all terms to lowercase and removing punctuation.

3.2 Results

On average, we found the length of a tile to be 4.47 for single-document summaries, compared with 2.33 for multi-document summaries. We discovered that 61 out of all 1667 hand-written single-document summary sentences exactly matched a sentence in the source document, however we did not find any sentences for which this was the case when examining multi-document summaries.

We also wanted to study how many sentences are fully tiled by phrases coming from exactly one sentence in the document corpus, and found that while no sentences from the multi-document summaries matched this criteria, 7.6% of sentences in the single-document summaries could be tiled in this manner. When trying to tile sentences with tiles coming from only one document sentence, we found that we could tile, on average, 93% of a single-document sentence in that manner, compared to an average of 36% of a multi-document sentence. This suggests that for multi-document summarization, we are not seeing any instances of what can be considered single-sentence compression. Table 1 summarizes the findings we have presented in this section.

	SingleDoc	MultiDoc
Average Tile Size (words)	4.47	2.33
Max Tile Size (words)	38	24
Exact sentence matches	3.7%	0%
Complete tiling from single sentence	7.6%	0%

Table 1. Comparison of Summary Tiling

Figure 1 shows the relative frequency with which a summary sentence is optimally tiled using tile-sizes up to 25 words in length in both the single and multi-document scenarios. The data shows that the relative frequency with which a single-document summary sentence is optimally tiled using n-grams containing 3 or more words is consistently higher compared to the multi-document case. Not shown on the histogram (due to insufficient readability) is that we found 379 tiles (of approximately 86,000) between 25 and 38 words long covering sentences from single-document summaries. No tiles longer than 24 words were found for multi-document summaries.

In order to test whether tile samples coming from tiling of single-document summaries and multi-document summaries are likely to have come from the same underlying population, we performed two one-tailed unpaired t-tests, in one instance assuming equal variances, and in the other case asssuming the variances were unequal. For these statistical significance tests, we randomly sampled 100 summary sentences from each task, and extracted the lengths of the n-grams found via minmal tiling. This resulted in the creation of a sample of 551 tiles for single-document sentences and 735 tiles for multi-document sentences.

For both tests (performed with α=0.05), the P-values were low enough (0.00033 and 0.000858, respectively) to be able to reject the null hypothesis that the average tile length coming from single-document summaries is the same as the average tile length found in multi-document summaries. We chose to use a one-tailed P-value because based on our experiments we already suspected that the single-document tiles had a larger mean.

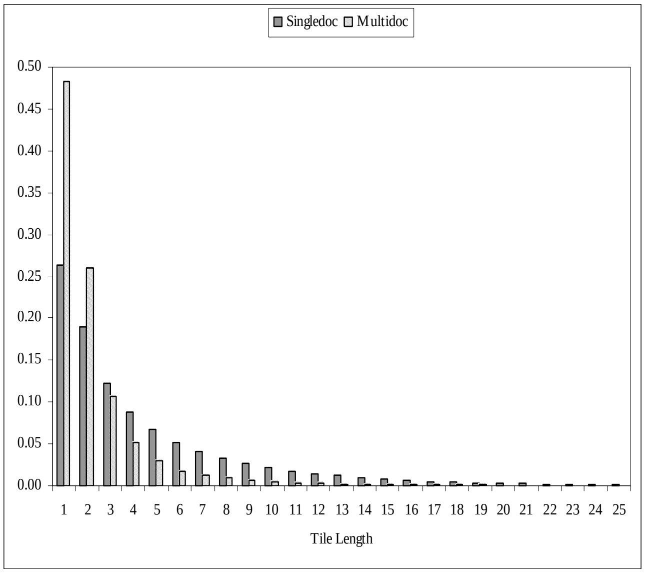

Figure 1. Comparison of Relative Frequencies of Optimal Tile Lengths

4 Conclusions and Future Work

Our experiments show that when writing multi-document summaries, human summarizers do not appear to be cutting and pasting phrases in an extractive fashion. On average, they are borrowing text around the bigram level, instead of extracting long sequences of words or full sentences as they tend to do when summarizing a single document. The extent to which human summarizers form extractive summaries during single and multi-document summarization was found to be different at a level which is statistically significant. These findings are additionally supported by the fact that automatic n-gram-based evaluation measures now being used to assess predominately extractive multi-document summarization systems correlate strongly with human judgments when restricted to the usage of unigrams and bigrams, but correlate weakly when longer n-grams are factored into the equation (Lin & Hovy, 2003). In the future, we wish to apply our method to other corpora, and to explore the extent to which different summarization goals, such as describing an event or providing a biography, affect the degree to which humans employ rewriting as opposed to extraction.

Despite the unique requirements for multi-document summarization, relatively few systems have crossed over into employing generation and reformulation (McKeown & Radev, 1995, Nenkova, et al. 2003). For the most part, summarization systems continue to be based on sentence extraction methods. Considering that humans appear to be generating summary text that differs widely from sentences in the original documents, we suspect that approaches which make use of generation and reformulation techniques may yield the most promise for multi-document summarization. We would like to empirically quantify to what extent current summarization systems reformulate text, by applying the techniques presented in this paper to system output.

Finally, the potential impact of our findings with respect to recent evaluation metrics should not be overlooked. Caution must be given when employing automatic evaluation metrics based on the overlap of n-grams between human references and system summaries. When reference summaries do not contain long n-grams drawn from the source documents, but are instead generated in the author's own words, the use of a large number of reference summaries becomes more critical.

Acknowledgements

The authors wish to thank Eric Ringger for providing tools used to construct the tiles, and Bob Moore for suggestions pertaining to our experiments.

References

Regina Barzilay, Noemie Elhadad, Kathleen McKeown. 2002. "Inferring Strategies for Sentence Ordering in Multidocument Summarization." JAIR, 17:35-55.

Jade Goldstein, Vibhu Mittal, Mark Kantrowitz and Jaime Carbonell, 2000. Multi-Document Summarization by Sentence Extraction. *In the Proceedings of the ANLP/NAACL Workshop on Automatic Summarization.* Seattle, WA

Hongyan Jing. 2002. Using Hidden Markov Modeling to Decompose Human-Written Summaries. Computational Linguistics 28(4): 527-543.

Chin-Yew Lin, and E.H. Hovy 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. *In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada.

Chin-Yew Lin and Eduard.H. Hovy. 2003. The Potential and Limitations of Sentence Extraction for Summarization. *In Proceedings of the Workshop on Automatic Summarization post-conference workshop of HLT-NAACL-2003.* Edmonton, Canada.

Daniel Marcu and Laurie Gerber. 2001. An Inquiry into the Nature of Multidocument Abstracts, Extracts, and Their Evaluation. *Proceedings of the NAACL-2001 Workshop on Automatic Summarization*

Inderjeet Mani and Eric Bloedorn. 1999. Summarizing similarities and differences among related documents. *Information Retrieval*, 1, pp. 35-67.

Inderjeet Mani and Mark Maybury (Eds.). 1999. *Advances in Automatic Text Summarization*. MIT Press, Cambridge MA.

Kathy McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of AAAI.*

Kathleen R. McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. 1995. In *Proceedings, 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74-82.

Ani Nenkova, Barry Schiffman, Andrew Schlaiker, Sasha Blair-Goldensohn, Regina Barzilay, Sergey Sigelman, Vasileios Hatzivassiloglou, and Kathleen McKeown. 2003. Columbia at the Document Understanding Conference. Document Understanding Conference 2003.

Paul Over. 2004. Personal communication.